Nonparametric and Semiparametric Survival Estimation in Two-Stage (Nested) Cohort Studies

Steven D. Mark

Abstract. Frequently in epidemiologic cohort studies the primary goal is to estimate the effect of exposures, V_i, on a time-to-event outcome, T_i, while adjusting for other covariates, J_i. When the cost of measuring V_i is disproportionate to the cost of J_i, it may be inefficient or infeasible to ascertain V_i on everyone. Cost may reflect financial cost, logistical cost, or health risks attendant upon obtaining V_i measurements from individuals. These considerations have given rise to two-stage sampling strategies: in stage-one J_i is observed on all members of a cohort; in stage-two a subgroup is selected for V_i measurement. For censored failure time data the most common two-stage designs are the case-cohort (CCH) and nested case-control designs (NCC). These focus on estimating the relative risk parameters in a Cox proportional hazards model. Rather than relative risk, our emphasis is on survival, or absolute risk. Though both CCH and NCC designs provide estimators of the cumulative hazards, and hence survivals, those estimators are biased if any cases are missing the V_i measurements. In this paper we present a class of nonparametric and semiparametric cumulative hazard estimators that are unbiased regardless of stage-two case-sampling fraction. We characterize the mathematical form of the efficient estimators; express the determinants of efficiency in terms of relative risks, survivals, and exposure prevalences; and describe how familiar subject matter considerations have practical implications for decisions regarding study design and analysis. We motivate this work with a data analysis of a two-stage study on *H. pylori* infection and gastric cardia cancer. Using simulations we demonstrate that differences in the efficiency of estimators accord with theory and can be substantial. We have written R and S-plus code that implements an approach to estimation that is conceptually simple and has desirable efficiency properties.

KEY WORDS: absolute risk; auxiliary covariate; cumulative hazard; efficiency; survival; two-stage studies; weighted estimating equations

Steven D. Mark, M.D., Sc.D.
Biostatistics Branch, Division of Cancer Epidemiology and Genetics
National Cancer Institute
6120 Executive Blvd. Room 8036 MSC 7244
Rockville, MD 20852-4910

301-402-9508 (voice)
301-402-0081 (fax)
sm7v@.nih.gov

1. Introduction

In epidemiologic cohort studies new exposures, which we call V_i, frequently become of interest after endpoints have already been recorded at follow-up time τ. Two-stage sampling designs are a common strategy for estimating the association of V_i with outcome in such cohorts. We focus on studies where the outcome is time to some event of interest, T_i. In the first stage of these studies, one observes a (possibly empty) set of covariates, A_i, and an outcome $\{X_i,\, \Delta_i\}$ for each of the n individuals. As usual, $X_i = min\,(T_i, C_i)$, C_i is a censoring time, $\Delta_i = I(X_i = T_i)$, and $I(\ \cdot\ )$ is the indicator function. Throughout this paper we assume censoring is independent and non-informative (see for example, Andersen, Borgan Gill, and Keiding, 1991). Consistent with epidemiologic parlance we call those with $\Delta_i = 1$ cases, and those with $\Delta_i = 0$, controls. W_i denotes the combined set of outcome and covariate data observed at the end of stage 1 (time τ).

$$W_i = \{X_i,\, \Delta_i,\, A_i\} \tag{1}$$

In the second stage of the study, using selection probabilities, $\pi_o(W_i)$, that depend only on W_i, a sub-sample of individuals is chosen for measurement of V_i. The motivation for sub-sampling is that V_i, which we subsequently refer to as the *exposures*, are in some sense, expensive, or difficult, to measure. Since the occurrence of cases is rare compared to that of controls, and case counts are the main determinants of the variance of the estimators, sampling rates are generally higher for cases than for controls. We define $R_i = 1$ if V_i is known for individual i; $R_i = 0$ otherwise. To control confounding, an investigator generally estimates the effect of V_i conditional on a set of *adjusting covariates*, J_i, $J_i \subseteq A_i$. We call variables in A_i that are not in J_i, *auxiliary variables*, and denote them Λ_i^{aux}.

When the outcome is time-to-event, the most common two-stage designs are the case-cohort (Prentice, 1986; Self and Prentice, 1988) and nested case-control designs (Lidell, McDonald, Thomas, 1977; Borgan, Goldstein, and Langholz, 1995). The primary focus of these designs has been estimating relative risks (*rr*) associated with covariates $Z_i = \{V_i, J_i\}$, when hazards are specified by a Cox proportional hazards model (CPH) such as (4). We recently reviewed these approaches and showed that the estimators and their variances can be written as a single set of estimating equations (Mark and Katki, 2001). In contrast, rather than estimating β_o, the focus in this paper is on the estimation of the conditional survivals, $S(\tau|z)$, where $z \in \mathcal{Z}$ (the support of Z_i).

The motivation for this work arises from two-stage studies we have conducted on a cohort in China with epidemic rates of gastric cardia stomach cancer (GCC). This cohort was selected from a well defined geographic population, and estimates of survival, or absolute risk, are of public health importance (Mark, Qiao, Dawsey, et al., 2000). To illustrate the issues and demonstrate an application of our procedures, we analyze data from our study on the association of *H.Pylori* (Hp) infection with incident GCC (Limburg PJ, Wang CQ, Mark SD, et al., 2000). In that study, V_i was the measurement of serum antibodies to Hp: V_i =1 if a subject had antibodies to Hp. A_i contained such information as age, sex, height, and weight. Since age was the only significant risk factor in A_i, in the analysis we present, J_i is an indicator variable, with $J_i = 1$ if a subject's age is greater than the median cohort age. We document the performance of estimators using simulations based on the structure of our current study on Hp and GCC which includes endpoints accrued over an additional ten years of follow-up.

Though our inferential focus is survival, the mathematical results we present are on the cumulative hazard scale, $\Lambda(\tau; z)$,

$$\Lambda(\tau; z) = \int_0^{\tau} \lambda(u|z)\, du \tag{2}$$

We obtain estimators of the conditional survival through the identity

$$S(\tau|z) = exp\,-(\,\Lambda(\tau; z)\,) \tag{3}$$

In 1994 Robins, Rotnitsky and Zhao (1994) (henceforth called RRZ), described the class of all two-stage estimators in terms of weighted estimating equations, and derived the mathematical form of the efficient member of the class. They focused on conditional mean models. Applying their results to time-to-event data, we describe the class of nonparametric and semiparametric cumulative hazard estimators. In nonparametric estimation no assumptions are made regarding the relationship between hazards at different levels of z. For the semiparametric model we assume the hazards are related by the Cox proportional hazards (CPH) model

$$\lambda(u|Z_i) = \lambda_o(u) exp\,(\beta_1^T V_i + \beta_2^T J_i); \tag{4}$$

where $Z_i = \{V_i^T, J_i^T\}^T$ and $\beta_o = \{\beta_1^T, \beta_2^T\}^T$ are conformable $p \times 1$ vectors of covariates and parameters. For simplicity, we assume that the Z_i are time invariant, and that, as expressed in (4), there is no V by J interaction. We refer to $\lambda_o(u)$ as the baseline hazard. Since our interest is in contrasting survivals of groups of individuals, in the body of the paper we assume Z_i has finite support of dimension k*. In the Hp data $k^* = 4$. In appendix D we give results on estimation in a completely general support space.

There are several practical advantages to the RRZ formulation. Both the case-cohort (CCH), and nested case-control (NCC) designs specify that cases be sampled with probability one. When V_i is not measured on all cases, the cumulative hazard estimators given in those proposals are biased (Mark and Katki, 2001). Epidemiology studies commonly require exposure measurements which are expensive and consume limited specimens. Consequently, designs with fractional cases-sampling have become increasingly frequent and attractive (Mark and Katki, 2001). In the Hp study, due to uncertainties with regard to the direction of the association and the prevalence of Hp infection, as well as a reluctance to use up the small quantities of available serum, we sampled approximately 25% of available GCC cases. The estimating equations we describe are weighted by the inverse of the sampling probability and accommodate any non-zero sampling rate. Even when the intent of an investigator is to measure V on all the cases, vagaries beyond investigator control seldom permit complete ascertainment (Mark et al, 2000; Mark and Katki, 2001). We describe the additional assumptions required for estimation when there is unplanned missingness in section 6.

Another feature distinguishing RRZ from CCH and NCC estimators, is that in RRZ estimators, individuals with unobserved V$_i$ contribute to estimation. In this paper we emphasis that the distinction between estimators within the RRZ class, and hence the differences in efficiency, are entirely due to variation in the extent to which information from subjects with R$_i = 0$ is utilized. We derive expressions for the efficient nonparametric, and the restricted-class efficient (defined in section 4) semiparametric estimators, in terms of aspects of the probability distribution of the data familiar to epidemiologists. This formulation has clear implications for the design and analysis of two-stage studies.

Finally, in this paper we emphasize a particular approach to estimation, which we call $\widehat{\pi}$-estimation. Formulation in terms of $\widehat{\pi}$-estimation provides a geometric representation, as well as a practical means of implementing, efficiency consideration. R and S-plus code that implements these $\widehat{\pi}$-estimators is available (Mark, 2003, Appendix F).

2. Full-Data Estimators and Influence Functions

We refer to studies in which V_i is observed for all n individuals as *full data studies,* and use $H_i = \{W_i, V_i\}$ to denote the fully observed data. In a sense made specific in section 4, RRZ proved that that all two-stage estimators and their corresponding influence functions can be expressed as *weighted versions with offset* of their full data counterparts. In this section we describe the full data estimators and influence functions for the nonparametric cumulative hazard, and for the semiparametric estimators of β_o and the baseline cumulative hazard (5) .

$$\Lambda_o(\tau, \beta_o) = \int_0^\tau \lambda_o(u)\ du \tag{5}$$

For the semiparametric model, the cumulative hazard at any covariate level z is $\Lambda(\tau; \beta_o z) = \Lambda_o(\tau, \beta_o) exp \beta_o^T z$, and is estimated in the obvious fashion. Its distribution is derived by the delta method, such as in Anderson et al. (1991). To indicate the $k^* \times 1$ vector of cumulative hazards at τ, we drop z from the arguments and write $\Lambda(\tau)$, or $\Lambda(\tau, \beta_o)$. In section 4 we provide corresponding results for two-stage estimators. We explicitly give results only in terms of estimation at the end of follow-up time τ; estimates at any other time t, $0 \leq t \leq \tau$, are obtained by substituting t for τ in the limit of integration of the estimators.

In full data studies the Nelson -Aalen estimator, $\widehat{\Lambda}(\tau, z)$, is the efficient nonparametric estimator of (2) (Anderson et al., 1991). The partial likelihood estimator, $\widehat{\beta}$, and the Breslow estimator, $\widehat{\Lambda}_o(\tau, \widehat{\beta})$ are the semiparametric efficient estimators of β_o (4) and the baseline cumulative hazard (5) (Anderson et al, 1991). Using standard counting process notation, we denote the event counting process $N_i(u)$, ($N_i(u) = 1$ iff $T_i \leq u$, and $T_i \leq C_i$), and the at risk process, $Y_i(u)$, ($Y_i(u) = 1$, iff $(C_i \wedge T_i) \leq u$). For individual i the hazard of T_i conditional on Z_i is $\lambda_i(u| Z_i) = Y_i(u) \times \lambda(u|Z_i)$. We assume the $\lambda(u|z)$ are non-negative, and the $\Lambda(\tau;$ $z)$ are finite. In the context of nonparametric estimation the above should be regarded as a multivariate counting process of dimension k^*. That is, we estimate the k^* within-stratum cumulative hazards, $\Lambda_h(\tau)$, where, for instance, the $h'th$ row of $N_i(u)$ is $N_{ih}(u) = 1$, iff $I(Z_i = h\,), T_i \leq u,$ and $T_i \leq C_i$. As is standard we define, $S^0(u) = \sum_{j=1}^{n} Y_i(u)$; $S^0(u, \widehat{\beta}) = \sum_{i=1}^{n} Y_i(u)\, exp\,(\widehat{\beta} Z_i\,)$; and $S^1(u, \widehat{\beta}\,) = \sum_{i=1}^{n} Y_i(u) Z_i\; exp\;(\widehat{\beta} Z_i\,)$. Under the usual regularity conditions (Anderson et al. ,1991), $n^{-1}\, S^j(u, \,\cdot\,) \overset{lim\,p}{\longrightarrow} E\big[S^j(u, \,\cdot\,)\big] = s^j(u, \,\cdot\,)$ for all three processes. $M_i(u\,)$ denotes the counting process martingale, $M_i(u) = N_i(u) - \Lambda_i(u)$.

The $k^* \times 1$ full data Nelson-Aalen estimator of $\Lambda(\tau)$ is (Anderson)

$$\widehat{\Lambda}(\tau) = \sum_{i=1}^{n} \int_0^{\tau} S^0(u)^{-1} dN_i(u) \tag{6}$$

Anderson et al. (1991) give the influence function expansion of $\widehat{\Lambda}\,(\tau\,)$ as

$$n^{\frac{1}{2}}\left(\widehat{\Lambda}(\tau) - \Lambda(\tau)\right) = n^{-\frac{1}{2}} \sum_{i=1}^{n} D_i^{F1} \; + o_p(1) \tag{7}$$

$$D_i^{F1} = \int_0^{\tau} \left[s^0(u)\right]^{-1} dM_i(u) \tag{8}$$

Newey (1990) showed that all nonparametric estimators have identical influence functions, and hence, are asymptotically equivalent. Thus (8) is the influence function for any nonparametric estimator of $\Lambda(\tau)$.

The class of full data estimators for the β_o in CPH model (4) (RRZ, 1994) are the $\widehat{\beta}(h)$'s that solve

$$\sum_{i=1}^{n} \int_0^{\tau} \left\{ h(Z_i, X_i) \;-\; S^1(s, \beta, h)\, S^0(s, \beta)^{-1} \right\} dN_i(s) = 0 \tag{9}$$

The choice of the function $h(Z_i, X_i)$ determines the efficiency of the estimator. For full data, the semiparametric efficiency bound is achieved by the partial likelihood estimator with $h(Z_i, X_i)$=Z_i. The full data influence function for the partial likelihood estimator is D_i^{F2}

$$D_i^{F2} = i^{-1} \int_0^{\tau} \left\{ Z_i - e(u, \beta_o) \,\right\} dM_i(u) \tag{10}$$

where $e(u,\beta_o) = s^1(u,\beta_o)s^0(u,\beta_o)^{-1}$, and i= $E\ \left(\int_0^\tau \left\{Z_i - e(u,\beta_o)\right\} dM_i(u)\right) \left(\int_0^\tau \left\{Z_i - e(u,\beta_o)\right\} dM_i(u)\right)^T$, the usual partial likelihood information. As in (7), the estimator $\widehat{\beta}$ can be expressed as the sum of its iid influence functions.

The Breslow estimator of the baseline cumulative hazard, $\Lambda_o(\tau,\beta)$, is given by

$$\widehat{\Lambda}_o(\tau,\widehat{\beta}) = \sum_{i=1}^{n} \int_0^\tau \left[S^0(u,\widehat{\beta})\right]^{-1} dN_i(u) \tag{11}$$

To obtain the influence function for (11) we write

$$\widehat{\Lambda}_o(\tau,\widehat{\beta}) - \Lambda_o(\tau,\beta_o) = \left\{\widehat{\Lambda}_o(\tau,\widehat{\beta}) - \ \widehat{\Lambda}_o(\tau,\beta_o)\right\} + \left\{\widehat{\Lambda}_o(\tau,\beta_o) - \Lambda_o(\tau,\beta_o)\right\} \tag{12}$$

Using a Taylor series expansion of $\widehat{\beta}$ around β_o as in Theorem VII 2.3 Anderson et al. (1991), we express the first term in the right hand side of (12) as

$$(\ \widehat{\beta} - \beta_o\ )^T \int_0^\tau e(u,\beta_o)\, \lambda_o(u)\, du + op(1)$$

Then replacing estimators in (12) with their influence functions, we have

$$n^{\frac{1}{2}}\left\{\ \widehat{\Lambda}_o(\tau,\widehat{\beta}) - \Lambda_o(\tau,\beta_o)\ \right\} = n^{-\frac{1}{2}} \sum D_i^{F3} + o_p(1) \tag{13}$$

$$D_i^{F3} = \int_0^\tau \left[s^0(u,\beta_o)\right]^{-1} dM_i(u)\ -\ D_i^{F2^T} \int_0^\tau e(u,\ \beta_o)\ d\Lambda_o(u,\beta_o)$$

We refer to the $D_i^{Fb}, b \in \{1,2,3\}$, as the *full data influence functions*. Like all influence functions, they are iid and have expectation 0. Hence the asymptotic variance of each estimator is $E\left[D_i^{Fb}\ D_i^{Fb^T}\right]$.

Though we explicitly present results for a non-stratified CPH model, the results of a stratified model, with strata generated from a discretization of A_i, can be described using the multivariate counting process.

3. Stage-Two Sampling Restrictions

For most of the paper we assume that conditional on $W_{i,}$ selection of individuals for measurement of V_i is independent with known, non-zero, probabilities, $\pi_o(W_i)$ that do no depend on V_i. That is

$$\pi(W_i) = Pr(\ R_i = 1 |\ W_i, V_i) = Pr(\ R_i = 1 |\ W_i) \tag{14}$$

In the usual parlance of missing data, restriction (14) is consistent with V_i being missing at random (MAR) (Rubin, 1976). As we frequently do for random variables, we drop the explicit argument of a function, and use the subscript i to indicate that it is a random variable. Thus we write $\pi_{i,o}$, where $\pi_{io} \equiv \pi_o(W_i)$. At the end of section 6 we extend the results to dependent sampling, and to missingness that is not entirely under investigator control.

Without loss of generality we specify the known sampling probabilities using the logistic model

$$logit\, \pi_o(W_i) = \psi_o^T\, h(W_i) \tag{15}$$

Here ψ_o and $h(W_i)$ are known, conformable, finite dimensional vectors of parameters and random variables, respectively. Clearly neither the parameterization nor the dimension of equation (15) are unique. For instance, if A_i contains only information on sex, and stage-two sampling depends only on case status, then two correctly specified models for (15) would be

$$logit\, \pi_o(W_i) = \psi_{o1} I(\Delta_i = 1) + \psi_{o2}\, I(\Delta_i = 0) \tag{16}$$

$$\begin{aligned} logit\, \pi_o(W_i) = \psi_{o1} I(\Delta_i = 1) + \psi_{o2}\, I(\Delta_i = 0) + \\ \psi_{o3}\, I(male) + \psi_{o4}\, I(female) \end{aligned} \tag{17}$$

Here $\psi_{o1} = logit\, Pr(R_i = 1|\Delta_i = 1)$; $\psi_{o2} = logit\, Pr(R_i = 1|\Delta_i = 0)$, and $\psi_{o3} = \psi_{o4} = 0$.

We define W_i^R to be the smallest set of linearly independent vectors such that (15) is true where size refers to the dimension of the column space spanned by the $h(W_i)$. In our example, the dimension of W_i^R is two. Correctly specified models are those with covariates W^l such that

$$W_i^l \geq W_i^R\ . \tag{18}$$

The inequality relation denotes that the span of W_i^l includes that of W_i^R. We consider models with equivalent spans to be identical, and restrict ourselves to covariate spaces where the W_i^l are linearly independent.

We denote the scores from any logistic model with covariates W_i^l as S_i^l,

$$S_i^l = (R_i - \pi_{io}) W_i^l \tag{19}$$

4.0 **Two-Stage Estimators and Influence Functions**

The two-stage risk set estimators, $\widetilde{S}^j(u, \cdot)$, are inverse probability weighted versions of the full data estimators. For instance, $\widetilde{S}^0_h(u) = \sum\limits_{j=1}^{n} \pi_{io}^{-1} R_i \, Y_{i,h}(u)$. Like their full data counterparts, their averages converge in probability to $s^j(u, \cdot)$ (Pugh, 1993; RRZ, 1994).

RRZ prove that two-stage estimators and their influence functions, can be expressed as weighted versions of the full data quantities with an "offset". Applying these results to nonparametric estimation establishes that all two-stage estimators of $\Lambda(\tau)$ are asymptotically equivalent to a member in the class of estimators, $\widetilde{\Lambda}(\tau,g_1)$, defined as

$$\widetilde{\Lambda}(\tau, g_1) = \sum_{i=1}^{n} \left\{ \int_0^{\tau} R_i \, \pi_{i,o}^{-1} \left(\widetilde{S}^0(u) \right)^{-1} dN_i(u) \, - \pi_{io}^{-1}(R_i - \pi_{io}) \; g_1(W_i) \right\} \qquad (20)$$

Here $g_1(W_i)$ is any $k^* \times$ 1vector of non-stochastic functions of W_i specified by the investigator. The corresponding influence functions are

$$D_i^1(g_1) = \; R_i \, \pi_i^{-1} D_i^{F1} - \pi_{io}^{-1}(R_i - \pi_{io}) g_1(W_i) \qquad (21)$$

Note that here, unlike the full data case where we have a single estimating equation (6) and influence function (8), there are a class of estimators and influence functions characterized by the "offset" $\pi_{io}^{-1}(R_i - \pi_{io}) \; g_1(W_i)$.

The class of two-stage semiparametric estimators of β_o are characterized by an $h(\cdot)$ function as well as the $p \times 1$ offset term, $\pi_{io}^{-1}(R_i - \pi_{io}) g_2(W_i)$ (RRZ,1994). Efficiency in estimation depends on the choice of both the $h(\cdot)$ and $g_2(\cdot)$ functions. The optimal function $h(\cdot)$ is a non-closed form integral equation that is a function of infinite dimensional parts of the survival and covariate distributions (RRZ, 1994). For reasons of practicality, we therefore follow a general recommendation of RRZ and restrict our estimators to the subclass that use the efficient full data $h(\cdot)$ function, $h(Z_i, X_i) = Z_i$. This subclass includes the CCH and NCC estimators. Hence, we consider estimators $\widetilde{\beta}(g_2)$ that solve

$$\sum_{i=1}^{n} \left(\int_0^{\tau} R_i \, \pi_i^{-1} \times \left\{ Z_i - \widetilde{S}^1(s, \beta) \, \widetilde{S}^0(s, \beta)^{-1} \right\} dN_i(s) - \pi_{io}^{-1}(R_i - \pi_{io}) g_2(W_i) \right) = 0 \qquad (22)$$

Due to this restriction, we refer to efficiency results for estimators of β_o, and $\Lambda_o(s,\beta_o)$ as restricted-class efficient estimators (RC-efficient). The influence functions for $\widetilde{\beta}(g_2)$ are

$$D_i^2(g_2) = \pi_{io}^{-1} R_i D_i^{F2}(\beta) \; - \pi_{io}^{-1}(R_i - \pi_{io}) g_2(W_i) \qquad (23)$$

The procedure for estimating the baseline cumulative hazard (5) is analogous to the full data case. First estimate $\widetilde{\beta}(g_2)$, then, with $g_{3^*}(W_i)$ any scalar function of W_i,

$$\widetilde{\Lambda}_o(\tau,\, \widetilde{\beta}(g_2), g_{3^*}) = \sum_{i=1}^{n} \pi_{io}^{-1} \left\{ R_i \int_0^\tau \left[\widetilde{S}^0(u,\, \widetilde{\beta}(g_2))\right]^{-1} dN_i(u) - (R_i - \pi_{io}) g_{3^*}(W_i) \right\}; \qquad (24)$$

Using an identical Taylor series expansion as above, the influence functions for (24) are

$$D_i^3(g_3) = \pi_{io}^{-1} R_i\, D_i^{F3} - \pi_{io}^{-1}(R_i - \pi_{io}) g_3(W_i); \quad g_{i,3} = g_{i,3^*} - g_{i,2} \int_0^\tau e(u,\, \beta_o)\, d\Lambda_o(u, \beta_o) \qquad (25)$$

As for the full data case, the asymptotic variances of the estimators are $E\left[D_i^b\, D_i^{b^T}\right]$.

Let $\widetilde{\Lambda}(\tau,\, \cdot\,)$ denote any non or semiparametric estimator of (2). We form estimates $\widetilde{S}(\tau, \cdot\, | v, j)$ of $S(\tau | v, j)$ by substituting $\widetilde{\Lambda}(\tau,\, \cdot\,)$ for $\Lambda(\tau)$ in (3). We provide consistent estimators of $E\left[D_i^{Fb}\, D_i^{Fb^T}\right]$, and the variances of the $\widetilde{S}(\tau, \cdot\, | v, j)$, are in appendix A.

5. The Efficient $\mathbf{g}_b(\,\cdot\,)$ for Estimators of $\Lambda(s)$ and β_o

We define simple true-π estimators (STP) as those in which $g_b = 0$ in 20,22,24 , and write the STP influence functions as, $D_i^b(\pi_o) \equiv \pi_{io}^{-1} R_i D_i^{Fb}$. We can then express the $D_i^b(g_b)$ as

$$D_i^b(g_b) = D_i^b(\pi_o) - \pi_{io}^{-1}(R_i - \pi_{io}) g_{i,b} \qquad (26)$$

From 20,22,and 24, it is apparent that differences between estimators in each class are entirely due to differences in the g_b. Thus, finding the minimum variance estimator is equivalent to finding the g_b that minimizes $E[D_i^b(g_b) D_i^{b^T}(g_b)]$. We call such a g_b, the efficient, or for the semiparametric models RC-efficient, g_b, and denote it by g_b^{eff}. For $b \in \{1, 2\}$, direct application of proposition 2.3 of RRZ establishes that $g_{i,b}^{eff} = E[\, D_i^{Fb} | W_i]$. For estimators of $\Lambda_o(s, \beta)$, which are a function of g_2 and g_{3^*}, we use RRZ 2.3 and show (Appendix B) that the minimum variance is obtained with $g_2 = 0,\ g_{3^*} = E[\, D_i^{Fb} | W_i]$, and that $g_3^{eff} = E\left[D_i^{F3} \middle| W\right]$.

RRZ prove that an equivalent representation of the influence functions in (26) is

$$D_i^b(W^l) = D_i^b(\pi_o) - q^b {S_i}^l \qquad (27)$$

Here q^b is any conformable matrix of constants, and S_i^l are scores (19) from correctly specified logistic models (15). In appendix C we use (27) to provide an alternative derivation for the efficiency results of RRZ 2.3. The proof relies on the following two characteristics of population least squares regression that are fundamental to understanding the $\widehat{\pi}$-estimating procedures, their efficiency properties, and their method of implementation: 1) for any given set of scores, S_i^l, the variance of $D_i^b(W^l)$ is minimized when q^b is the projection operator, P^{bl}, of $D_i^b(\pi_{i,o})$ on S_i^l.

$$P^{bl} = E[D_i^b(\pi_o)S^{l^T}]E[S_i^l S_i^{l^T}]^{-1} \tag{28}$$

2) Since $D_i^b(\pi_{i,o}) - P^{bl}S_i^l$ is the residual from a projection, the variance is non-decreasing in the dimension of W_i^l. In appendix C we show that the minimum variance is reached when

$$P^{bl}\,S_i^l = \pi_{io}^{-1}\,(R_i - \pi_{io})\,E\left[D_i^{Fb}\,|W_i\right] \tag{29}$$

and that (29) (Appendix C, Result 2) is true for logistic model (30)

$$logit\,\pi_o(W_i) = \psi_1^T\,h(W_i) + \psi_2'\,W_{i,b}^{eff}; \quad W_{i,b}^{eff} = \pi_{io}{}^{-1}\,E\left[D_i^{Fb}\,\middle|\,W_i\right] \tag{30}$$

6. $\widehat{\pi}$-Estimators

We define $\widehat{\pi}$-estimators to be the solution to estimating equations 20,22, and 24 when $g_{i,b} = 0$ and the known sampling probabilities, $\pi_{i,o}$, are replaced with predicted sampling probabilities, $\widehat{\pi}_i(W^l)$. Specifically, the $\widehat{\pi}_i(W^l)$ are formed by replacing ψ_o in (15) with maximum likelihood estimates, $\widehat{\psi}$. RRZ (proposition 6.1) show that $\widehat{\pi}$ -estimators are consistent, asymptotically normal, with influence function

$$D^b{}_i(\widehat{\pi}(W^l)\,) = D_i^b(\pi_{i,o}) - P^{bl}S_i^l \tag{31}$$

It immediately follows that the variance of any $\widehat{\pi}(W^l)$-estimator is less than or equal to the variance of the STP estimator; and, for $W_i^m > W_i^l$, the variance of the $\widehat{\pi}(W_i^m)$ estimator is less than or equal to the variance of the $\widehat{\pi}(W_i^l)$ estimator. In Result 1 appendix C, we show that if a $\widehat{\pi}$-estimators is based on a logistic model saturated in W_i^f, such as model (17) , then $P^{bf}S_i^f = \pi_{io}^{-1}(R_i - \pi_{io})E\left[D_i^{Fb}\middle|W_i^f\right]$. Mark (2003, Appendix F) provides code for implementing $\widehat{\pi}$-estimators using any logistic model (15).

One feature of $\widehat{\pi}$-estimation is that it is the "natural" estimating procedure when the requirements that sampling is independent and with known probabilities are relaxed. In general, the dependent sampling we consider is characterized as follows: partition the observed W_i into a finite number of strata; select a fixed number of cases and controls from each stratum. If we let W_i^f be the saturated column space of indicator variables generated by that partition, then we can use any $\widehat{\pi}$-estimator with $W_i^l \geq W_i^f$ (RRZ, lemma 6.2). Such dependent sampling commonly occurs. For example, in the Hp study we sampled a fixed number of cases and controls. NCC risk set sampling is by design dependent. Mark (2003, Appendix E) provides $\widehat{\pi}$-estimators for both the NCC and CCH sampling schemes.

We have so far assumed that sampling probabilities, π_{io}, are entirely under investigator control. Suppose, however, some missingness occurred by chance. Under the assumption that this missingness was also MAR (14), and that, rather than knowing ψ_o, the investigator can specify a correct model such that $logit\ \pi_{io} = \psi^* W_i^l$ for some ψ^*, then the estimator $\widehat{\pi}(W_i^l)$ has influence function given by given by (31) (RRZ, proposition 6.2). For instance, in our study, due largely to mishaps in serum storage, approximately 10% of individuals had no serum on which Hp antibodies could be measured. Given the nature of the events causing the missingness, we believe that missingness was related to neither W_i or V_i. Hence, any $\widehat{\pi}$-estimator with $W_i^l \geq W_i^R$ would be consistent.

Section 7: Analyses of the Hp Data Using the $\widehat{\pi}(\Delta,J)$-estimator

Though Hp infection is a well established risk factor for gastric cancers arising outside of the cardia of the stomach (Helicobacter and Cancer Collaborative Group, 2001), the association with gastric cancers that arise in the cardia region (the proximal 2-3 centimeters of the stomach) is less established. Prior to our study, only a few small studies (case sizes ranging from 4 to 12), examined the Hp-GCC association. The consensus from these studies (Helicobacter and Cancer Collaborative Group, 2001; Dawsey, Mark, Taylor, et al., 2002), all conducted on Western populations, was that Hp was "protective" for GCC, with *rr* $\approx$ 0.5. Various mechanistic hypothesis have been advanced to account for the opposite association of Hp on GNC and GCC (Blaser, 1999).

In our study (Limburg et al. 2001) we sampled approximately 25% of GCC cases (100 cases) and 7% of controls (200 controls) that occurred in the cohort of 30,000 by 5.25 years of follow-up. We found an Hp prevalence (Hp$^+$) of approximately 65%, and a *rr* of approximately two for Hp$^+$ individuals. The only other major independent risk factor for GCC in this population was age: age greater than the cohort median age increased GCC risk by a factor of 3.5.

Table 1 contains estimates of covariate specific survivals at $\tau = 5.25$ years based on the CPH model (4) with V (Hp) and J (age) indicator variables as defined in the introduction. We used the $\widehat{\pi}$-estimator based on logistic model (17). Throughout this paper we denote this estimator as $\widehat{\pi}(\Delta, J)$. At each level of age, the Hp$^+$ group had lower survivals than the Hp$^-$ group. Within levels of Hp exposure, survival was higher in the younger group. Using the population age distribution for standardization (see A.4 for definitions and formulae), we estimated that Hp$^+$ individuals had 1.8% more cases (95% CI, 0.02-2.15) of GCC than the Hp$^-$ individuals in the 5.25 years of follow-up. $(INSERT\ \ TABLE\,1\ \ HERE\,)$

8. Implications of Efficiency for Study Design and Analysis

8.1 The general case

The optimal $g(\,\cdot\,)$ function, $E\left[D_i^{Fb}\middle|W\right]$, is a function of unknown parameters. RRZ proposition 2.4 established that g_b^{eff} can be replaced by a consistent estimator, $\widehat{g}_b^{eff}$, without changing the asymptotic distribution of the estimator. That is, an estimator using $\widehat{g}_b^{eff}$ achieves the nonparametric efficiency, or semiparametric RC-efficiency, bound. When g_b^{eff} can be consistently estimated from the data and model assumptions, we say the efficient estimator is identified. If not, then the variance of the efficient influence function represents an unknown lower bound that no estimator is guaranteed to achieve. It is immediately clear that unless X_i is a deterministic function of $\{\Delta_i, A_i\}$, $E\left[D_i^{Fb}\middle|W\right] \neq E\left[D_i^{Fb}\middle|\Delta_i, A_i\right]$, and efficient estimation requires X_i in the conditioning event. In the remainder of this section we approach the task of conditioning on X_i, by re-expressing g_b^{eff} in terms of relative risks, survivals, and covariate distributions. We discuss conditions under which each of these can be consistently estimated, and examine the implications for study design and analysis.

We re-express $g_{i,b}^{eff}$ as

$$g_{i,b}^{eff}{=}EE\left[D_i^{Fb}\middle|W_i, V_i\right] = \int_{\mathcal{V}} D_i^{Fb}\Big(W_i,\, v\Big)\, Pr(v|W_i)dv \tag{32}$$

In the design stage, a crucial consideration is what, if any, auxiliary variables should be measured. From (32) it is clear that for Λ_i^{aux} to contain the "optimal set" of covariates, it is sufficient that for any larger set, $\Lambda_i^{aux+} > \Lambda_i^{aux}$,

$$Pr(v|X_i, \Delta_i, J_i, \Lambda_i^{aux}\,) = Pr(v|\,X_i, \Delta_i, J_i, \Lambda_i^{aux+}) \tag{33}$$

That is, we should collect all auxiliary information which provides additional knowledge about the distribution of the incompletely measured covariates V_i at any time on study. To further examine the determinants of $Pr(v|W_i)$, we reparameterize in terms of the time dependent exposure odds

$$K_{i,v^\dagger} \equiv K_{v^\dagger}(W_i) = Pr(V_i = v^\dagger|\Delta_i, X_i, A_i)\Big/Pr(V_i = v^1|\Delta_i, X_i, A_i) \tag{34}$$

where $v^1$is some chosen reference level in $\mathcal{V}$, and $v^\dagger \in \mathcal{V}$. Using Bayes' theorem and a non-informative censoring assumption we show (Appendix D) that

$$K_{i,v^\dagger} = rr\,(X_i|\,v^\dagger, A_i)^{\Delta_i} \times S(X_i|v^\dagger A_i)\,\Big/S(X_i|v^1 A_i)\,\times Pr(v^\dagger|A_i)\Big/Pr(v^1|A_i) \tag{35}$$

Here $rr(X_i|\,v^\dagger, A_i)$ and $S(X_i|v,A_i)$ are the relative risks and survival probabilities at X_i conditional on $\{V_i, A_i\}$, rather than $\{V_i, J_i\}$. By (35), (33) is true if, for all times u

$$S(u|V_i, J_i, \Lambda_i^{aux}) = S(u|V_i, J_i, \Lambda_i^{aux+}) \tag{36}$$

and

$$Pr(V_i|J_i, \Lambda_i^{aux}) = Pr(V_{\,i}|J_i, \Lambda_i^{aux+}) \tag{37}$$

Epidemiologists refer to (36) as Λ_i^{aux}containing all *independent predictors of outcome;* and (37) as Λ_i^{aux} containing all *independent predictors of exposure*.

The requirements for efficient analysis are conceptually and mathematically equivalent to those in the design stage. That is, to estimate g_b^{eff}, we need only include in the conditioning event that subset of Λ^{aux} that contains the independent predictors of outcome and exposure. Though for any given Λ_i^{aux} it is impossible to know with certainty whether (36) or (37) are true, these are the exact considerations required to control confounding. Consequently, in the analysis stage epidemiologists generally try to choose J_i as the subset of A_i such that (36) and (37) are "approximately" true when Λ_i^{aux} is removed from the conditioning event on the left hand side. If successful in selecting J_i so that it contains all the *independent predictors of outcome*, (35) becomes

$$K_{i,v^\dagger} = rr\,(X_i|\,v^\dagger, J_i)^{\Delta_i} \times S(X_i|v^\dagger, J_i)\Big/S(X_i|v^1, J_i) \times Pr(v^\dagger|A_i)\Big/Pr(v^1|A_i) \tag{38}$$

8.2 Efficiency when J contains all the independent risk factors

In this section we assume that J_i contains all the independent predictors of outcome in A_i, so that (38) is true. From (32) it is clear that we can estimate $g_{i,b}^{eff}$, if we can estimate each of the terms in $K_{i,v^\dagger}$. For both the non and semiparametric failure time models, the second and third terms can be estimated by $\widetilde{S}(X_i|v^\dagger J_i)$ and $\widehat{P}(v^\dagger|A_i)$, where $\widehat{P}(v^\dagger|A_i)$ is the empirical average of V within levels of A. For the semiparametric model, $rr(u|Z_i)$ can be estimated by $r\widetilde{r}\,(u|Z_i) = exp\,\widetilde{\beta}^T Z_i$. Here the $\widetilde{S}(X_i|v^\dagger J_i)$ and $\widetilde{\beta}$ come from estimates based on any g_b. Hence the semiparametric RC-efficient estimators of β_o and $\Lambda_o(\tau, \beta)\, exp\beta_o^T Z$ are identified. In contrast, the nonparametric model provides no obvious estimator of $rr(u|Z_i)$. If k^* were small, and the number of cases large, one could theoretically use kernel smooths to estimate hazards, and hence *rr*'s. We do not explore this possibility further. Instead, in section 9 we propose several *locally efficient estimators (LE-estimators).* LE-estimators approximate g_b^{eff} by making assumptions about $rr(u|Z_i)$. We denote the resultant approximations by $\widehat{g}_b^{?eff}$. If the assumptions about the *rr*'s are correct, then $\widehat{g}_b^{?eff} \overset{lim\,p}{\longrightarrow} g_b^{eff}$, and the LE-estimators are efficient. Regardless of the truth of the assumptions, the proposed LE-estimators are consistent.

9. Simulations of STP, $\widehat{\pi}$, RC-efficient and Locally Efficient Estimators

All simulations are based on the following covariate distribution: $Pr(J1) = 0.5$, $Pr(V1) = 0.65$; $Pr(V1|J1) = 0.85$ (here, and in what follows, we use the abbreviated notation $J1$ for $J = 1$). T_i was specified by a CPH model with exponential baseline hazard. The magnitudes for the baseline hazard, and the exponential hazard for the independent censoring times, were chosen to produce approximately 1000 expected cases in a cohort of size n=6600 by time τ. This is the approximate number of cases that have occurred through the latest endpoint assessment, τ= 15 years. For the semiparametric models we simulated under the two covariate CPH model (4) with $\beta_1 = ln\,2$ $(rr_v = 2)$, $\beta_2 = ln\,3\,(rr_j = 3)$, and estimated $S(\tau|v, j)$. For the nonparametric simulations the data were generated by a one-covariate CPH model with $\beta_1 = ln\,2$ $(rr_v = 2)$; we estimated $S(\tau|v)$. Stage 2 sampling was always binomial, depending only on case status (16). Fifteen per cent of controls and 25% of cases were sampled, with a resultant control-to-case ratio of approximately 3:1. Each of the simulation results represents the average of 2000 realizations. Since, as evident from the tables, all of the survival estimators (and estimators of β_o, data not shown) were unbiased and had confidence intervals that covered near the stated rates, we focus the discussion on *relative efficiency (RE)*. RE is defined as the ratio (times 100) of the variance of a given estimator to the variance of the STP estimator. The smaller the RE, the greater the efficiency.

Table 2 contrasts the STP, RC-efficient, and $\widehat{\pi}(\Delta, J)$ semiparametric estimators of $S(\tau|v0j0)$ and $S(\tau|v1j1)$. Both the RC-efficient and the $\widehat{\pi}(\Delta, J)$ estimators are substantially more efficient than the STP estimator: approximately 45% more efficient in estimating $S(\tau|v0j0)$, and 70% more efficient in estimating $S(\tau|v1j1)$. The greater magnitude of the gains for $S(\tau|v1j1)$ reflects the fact that, in general, efficiency differences are due to the differential extraction of information from cases with unmeasured V_i. In this simulation approximately 8% of cases had V0J0, whereas 71% had V1J1. The differences in efficiency as a function of covariate values disappear when the simulations are set to produce equal number of cases in each covariate level (data not shown). Though for both covariate levels the RC-efficient are more efficient than the $\widehat{\pi}(\Delta, J)$-estimators, these differences are small. Since for the saturated $\widehat{\pi}(\Delta, J)$-estimator g_b=$E\left[D_i^{Fb}\middle|J_i, \Delta_i\right]$ (Result 1, Appendix C), the slight advantage of the RC-efficient reflects the fact that little is gained by adding the actual observed time, X_i, to the conditioning events. *(INSERT TABLE 2 HERE)*

Table 3 contains results for nonparametric estimators of $S(\tau|v)$ when J_i is an auxiliary covariate rather than a risk factor. For example, J_i might be a surrogate for V_i, such as evidence of gastric inflammation found on a biopsies obtained at the beginning of the study. These simulations reveal two important features of estimation. First, they demonstrate the potential for gaining efficiency by utilizing auxiliary information: the $\widehat{\pi}(\Delta, J)$-estimator (logistic model 17) is more efficient than the $\widehat{\pi}(\Delta)$-estimator (logistic model 16). In simulations (not shown) where V and J are independent, the efficiency of the $\widehat{\pi}(\Delta, J)$-estimator is identical to that of $\widehat{\pi}(\Delta)$-estimator. Second, the contrast in the performance of the two different locally efficient estimating procedures illustrates some noteworthy properties of LE-estimators. Each of the corresponding named *simple* (*SLE*) and *insured* (*ILE*) *local efficient* estimators use identical estimates, $\widehat{g}_1^{?eff}$, of g_1. However SLE-estimates are produced by setting $g_1 = \widehat{g}_1^{eff}$ in (20), whereas ILE-estimates are $\widehat{\pi}$-estimates based on prediction model (30) with $g_1 = \widehat{g}_1^{eff}$. By construction, ILE-estimators must be at least as efficient as $\widehat{\pi}(\Delta, J)$-estimators, even when $\widehat{g}_1^{?eff}$ is based on a misspecified $rr(X_i|v1)$. SLE's do not share this property. For example, the $\widehat{g}_1^{?eff}$ of the *SLE and ILE correct-estimators* is based on a correctly specified models for $rr(X_i|v1)$. Specifically, we assumed exponential hazards within each V level; estimated the hazards by dividing the number of observed cases by total person-time; and estimated $rr(X_i|v1)$ as a ratio of the hazards. Both the SLE and ILE correct estimators attain the nonparametric efficiency bound. In contrast the SLE and ILE *prior* and *null* estimators use misspecified $rr(X_i|v1)$'s. The *prior* estimators set $rr(X_i|v1) =$ 0.5, the pooled estimate of *rr* from the prior studies. The *null* estimators set $rr(X_i|v1) = 1$; these would be the efficient estimator under the null hypothesis. Table 3 shows that for estimators of $S(\tau|v0)$ the SLE-prior estimator is less efficient than the $\widehat{\pi}(\Delta, J)$-estimator. In simulations with V and J independent (data not shown), the SLE-prior has a variance 9% greater than even the STP-estimator. Thus the RE's of the SLE estimators are not bounded above by 1. In contrast, in the Table 3 simulations the ILE prior is more efficient that the $\widehat{\pi}(\Delta, J)$-estimator. Under independence the efficiencies are nearly identical. Note that all locally efficient estimators, misspecified or not, are unbiased and have confidence intervals that cover at the stated rate. *(INSERT TABLE 3 HERE)*

10. Discussion

Two-stage studies are commonly used in epidemiology as a resource-effective means of estimating the association of disease with exposures whose measurements consume a substrate of limited quantity. When estimating survival, the procedures proposed by the case-cohort and nested case-control designs are biased if cases are missing exposure measurements. By chance alone it is rarely possible to make measurements on all cases of a cohort. Applying results of RRZ, we derived a class of nonparametric estimators, and a class of semiparametric estimators, that provide unbiased estimates of cumulative hazards and survivals when cases are missing covariate data. We used a semiparametric estimator to analyze data from a study we conducted on the association of *H. pylori* infection and gastric cardia cancer in which only twenty-five percent of available cases were sampled. We found significant differences in age-standardized survivals between subjects with and without Hp infection.

Differences in efficiency between estimators results from differences in their utilization of information from subjects with no stage-two measurements. The standard NCC and CCH designs exclude those individuals from estimation. More efficient estimators use the data observed in stage-one to provide information on the V_i exposures not observed in stage-two. We expressed the optimal estimators in terms of the familiar quantities of relative risks, survivals, and exposure prevalences. Based on those expressions we described various estimating strategies that allow investigators to incorporate knowledge, estimates, or hypothesis about those quantities in a manner which can increase efficiency without sacrificing consistency. The insured local estimating procedures we proposed provide the additional guarantee that even if the incorporated knowledge is incorrect, efficiency will not decrease. We further showed that the subject matter considerations required to control confounding in observational studies are identical to those required when considering efficiency: investigators should measure all covariates thought to be independent predictors of either exposure or disease. Through simulations we demonstrated that the variation in efficiency between estimators within a class is of practical consequence. We emphasized a general approach to estimation, $\widehat{\pi}$-estimation, which allows investigators a flexible approach to specifying estimators with desirable efficiency properties. We have written and documented computer code in S-plus and R for these $\widehat{\pi}$-estimators (Mark, 2003, Appendix F) . These allow estimation of survivals and relative risks in a completely general covariate space.

References

P.K. Andersen., O. Borgan., R.D. Gill, and N. Keiding, "Statistical Models Based on Counting Processes", Springer-Verlag, New York, 1991.

M.J. Blaser, "Hypothesis: The changing relationship of Helicobacter pylori and humans: implications for health and disease," *Journal of Infectious Diseases*, vol. 179 pp. 1523-1530, 1999.

W.J. Blot, J.Y. Li, P.R. Taylor, W. Guo, S. Dawsey, G.Q. Wang, C.S. Yang, S.F. Zheng, M. Gail, G.Y. Li, Y. Yu, B.Q. Liu, J. Tangera, Y.H. Sun, F.S. Lie, J.F. Fraumeni, Y.H. Zhang, B. Li, "Nutrition intervention trials in Linxian, China: supplementation with specific vitamin/mineral combinations, cancer incidence, and disease-specific mortality in the general population," *Journal of the National Cancer Institute*, vol. 85 pp. 1483-1492, 1993.

O. Borgan, L. Goldstein., and B. Langholz., "Methods for the analysis of sampled cohort data in the Cox proportional hazards model," *The Annals of Statistics*, vol. 23 pp. 1749-1778, 1995.

S.M. Dawsey, S.D. Mark, P.R. Taylor, and P.J. Limburg, "Gastric Cancer and H Pylori." *Gut*, 51, 457-458, 2002.

Helicobacter and Cancer Collaborative Group. "Gastric cancer and Helicobacter Pylori: a combined analysis of 12 case-control studies nested within prospective cohorts," *Gut*, vol. 3 pp. 347-353, 2001.

F.D.K. Lidell, J.C. McDonald, and D.C. Thomas, "Methods for cohort analysis: appraisal by application to asbestos mining (with discussion)," *Journal of the Royal Statistical Society Ser, A*, vol. 140 pp. 469-490, 1977.

P.J. Limburg, C.Q. Wang, S.D. Mark, Y.L. Qiao, G.I Perez-Perez, M.J. Blaser, P.R. Taylor, Z.W. Dong, and S.M. Dawsey, "Helicobacter Pylori Seropositivity: Association with Increased Gastric Cardia and Non-Cardia Cancer Risks in Linxian, China," *Journal of the National Cancer Institute*, vol. 93 pp. 226-233, 2001.

S.D. Mark, Y.L. Qiao, S.M. Dawsey, H. Katki, E.W. Gunter, W. Yan-Ping, J.F. Fraumeni, W.J. Blot, Z.W. Dong, and P.R. Taylor, "Higher serum selenium is associated with lower esophageal and gastric cardia cancer rates," *Journal of the National Cancer Institute*, vol. 92 pp. 1753-1763, 2000.

S.D. Mark., and H. Katki, "Influence function based variance estimation and missing data issues in case-cohort studies," *Lifetime Data Analysis*, vol. 7 pp. 329-342, 2001.

S.D. Mark, "Nonparametric and semiparametric survival estimation in two-stage (nested) Cohort Studies." *2003 Proceedings of the American Statistical Association, Statistics in Epidemiology Section [CD-ROM]*, Alexandria, VA: American Statistical Association. In press.

W.K. Newey, "Semiparametric Efficiency Bounds", *Journal of Applied Econometrics*, vol. 5 pp. 99-135, 1990.

R.L. Prentice, "A case-cohort design for epidemiologic cohort studies and disease prevention trials," *Biometrika*, vol. 73 pp. 1-11, 1986.

M.G. Pugh, *Inference in the Cox Proportional Hazards Model with Missing Covariate Data*, thesis, Harvard School of Public Health: Boston, MA, 1993.

J.M. Robins, A. Rotnitsky, and L.P. Zhao, "Estimation of regression coefficients when some regressors are not always observed," *Journal of the American Statistical Association*, vol. 89 pp. 846-866, 1994.

D.B. Rubin, "Inference and Missing Data," *Biometrika*, vol. 63 pp. 581-592, 1976.

S.G Self., and R.L Prentice, "Asymptotic distribution theory and efficiency results for case-cohort studies," *The Annals of Statistics*, vol. 16 pp. 64-81, 1988.

Appendix A.

In this appendix, we provide consistent estimators of the asymptotic variances for the two-stage estimators of cumulative hazards, relative risks, and survivals. When we can do so without confusion, and to indicate that any consistent estimator of a parameter will suffice, we drop the arguments g_b. For instance, we write $\widetilde{\Lambda}(\tau)$ for $\widetilde{\Lambda}(\tau,g_1)$. We further define $d\widetilde{M}_i(u) = dN_i(u) - Y_i(u)\, d\widetilde{\Lambda}(u)$; $d\widetilde{M}_i(u, \beta) = dN_i\,(u) - Y_i(u)\, d\widetilde{\Lambda}_o(\, u, \widetilde{\beta}\,) exp\widetilde{\beta} Z_i$; $\widetilde{s}^j(u,\ \cdot\,) = n^{-1} \widetilde{S}^j(u,\ \cdot\,)$; $\widetilde{e}(u, \beta) = \widetilde{s}^1(u, \beta)\widetilde{s}^0(u, \beta_o)^{-1}$; $\widetilde{i} = n^{-1}\sum_{i=1}^{n} \pi_{i,o}^{-1} R_i\, \Delta_i \Big(\, Z_i - \widetilde{E}(X_i, \beta)\, \Big)\Big(Z_i - \widetilde{E}(X_i, \beta)\, \Big)^T$;

A.1. Estimating $\boldsymbol{D_i^b(g_b)}$ (26), and $\boldsymbol{D_i^b(\widehat{\pi}(W^l))}$ (31)

Estimators $\widetilde{D}_i^b(g_b)$ of $D_i^b(g_b)$ are formed by the obvious substitutions for $s^j(u,\ \cdot\,)$, $dM_i(u,\ \cdot\,)$, and $\widetilde{e}(u, \beta)$ in 21, 23, 25. The weights $\pi_{i,o}$ can be replaced by any consistent estimate, $\widehat{\pi}$. For $\widehat{\pi}$-estimators, $\widetilde{D}_i^1\,(\,\widehat{\pi}(W^l)\,)$ and $\widetilde{D}_i^2(\widehat{\pi}(W^l)\,)$ are formed by estimating P^{bl} (28) by the vector of regression parameters from an ordinary least squares regression of $\widetilde{D}_i^b(\pi_{io}\,)$ on the scores $\widetilde{S}_i^l$. Letting $g_{i,2}(\widehat{\pi}) = \pi_{i,o} P^{2l} W_i^l$, we express $D_i^3(\widehat{\pi}(W^l))$ as

$$D_i^3(\widehat{\pi}(W^l)) \;\equiv D_i^3(\, g_2(\widehat{\pi}), g_{3^*} = 0\,) - E[D_i^3(\, g_2\,(\widehat{\pi}), g_{3^*} = 0) S^{1^T}] E[S_i^1\, S_i^{l^T}]^{-1}\; S_i^l$$

and form the estimator, $\widetilde{D}_i^3(\widehat{\pi}(W^l))$, using the ordinary least squares regression as above.

A.2 Estimating the asymptotic variance of $\boldsymbol{\widetilde{\Lambda}(\tau)}$, and $\{\boldsymbol{\Lambda_o}(\tau, \boldsymbol{\beta}), \widetilde{\boldsymbol{\beta}}^T\}^T$

Let V_1 and V_a be the variances of $\widetilde{\Lambda}(\tau)$ and $\{\widetilde{\boldsymbol{\Lambda}}_{\boldsymbol{o}}(\tau,\widetilde{\boldsymbol{\beta}}),\ \widetilde{\boldsymbol{\beta}}^T\}^T$ respectively. Let $\widetilde{D}_i^a = \{, \widetilde{D}_i^3, \widetilde{D}_i^{2^T}\}^T$. Consistent estimates of the asymptotic variance are $\widetilde{V}_1 = n^{-1} \sum \widetilde{D}_i^1 \widetilde{D}_i^{1^T}$ and $\widetilde{V}_a = n^{-1} \sum \widetilde{D}_i^a\, \widetilde{D}_i^{aT}$.

A.3 Estimating the asymptotic variance of $\widetilde{S}(\tau | v, j)$.

Let $\overrightarrow{\widetilde{S}}(\tau)$ and $\overrightarrow{\widetilde{S}}(\tau,\beta)$ be the $k^* \times 1$ vector of nonparametric and semiparametric survival estimators, with row h entry $\widetilde{S}(\tau\, | z = h)$ and $\widetilde{S}(\tau,\ \beta |\ z =\ h)$, $h \in \mathcal{Z}$. Let V_{s1} and V_{s2} be the corresponding $k^* \times k^*$ variance matrices for $\overrightarrow{\widetilde{S}}(\tau)$ and $\overrightarrow{\widetilde{S}}(\tau,\beta)$. Define G as the $k^* \times k^*$

diagonal matrix with $\widetilde{S}(\tau|z=h)$ in the h'th row h'th column. Then $\widetilde{V}_{s1} = G\widetilde{V}_1 G_1$ is a consistent estimate of V_{s1}. Each $h \in \mathcal{Z}$ corresponds to a unique $p \times 1$ covariate vector, z_h. Let L_h be the $1 \times (p+2)$ vector, $L_h = \widetilde{S}(\tau, \beta|h)\ exp(\widetilde{\beta}^T z_h) \times \{1,\ \widetilde{\Lambda}_o(\tau,\beta) \times z_h^T\}$. Let L be the $k^* \times (p+1)$ matrix with $h'th$ row L_h. Then $\widetilde{V}_{s2} = L\widetilde{V}_a L^T$ is a consistent estimator of V_{s2}.

A .4 Standardized survival and standardized risk differences.

Consistent with common usage we define *standardized survival*, $S^s(\tau|v)$, to be the weighted sum of covariate specific survivals, with known weights, $w(j^*)$, which sum to 1. That is, $S^s(\tau|v) = \sum_{\mathcal{J}} S(\tau|v, j^*)\, w(j^*)$. Let v^*, j^* be the number of levels of V and J respectively. Arrange $\overrightarrow{\widetilde{S}}(\tau, \cdot\ |v,j)$ in v^* blocks of length j^*, in order of increasing index. Let W_j^T be the $1 \times j^*$ matrix of weights w_j; I_{v^*} the $v^* \times v^*$ identity matrix; and $C_w = W_j^T \otimes I_{v^*}$ where $\otimes$ denotes the Kronecker product. Then $\overrightarrow{\widetilde{S}}^s(\tau|v) = C_w\, \overrightarrow{\widetilde{S}}(\tau \cdot\ |v,j)$ with variance estimated by, for instance, $C_w\, \widetilde{V}_{s1}\, C_w^T$. Estimates of standardized risk differences , $\widetilde{R}d(\tau)$, are simple contrasts of the $\widetilde{S}^s(\tau|v)$.

Appendix B

In this appendix we prove that the variance of (24) is minimized when $g_2 = 0,\ g_{3^*} = E[D_i^{F3}|W_i]$. Let $C_{i1}(g_2) = \pi_{io}^{-1} R_i\ D_i^{F3} + i^{-1}\pi_{io}^{-1}(R_i - \pi_{io})\, g_{i2}\, k_1;\ k_1 = \int_0^\tau e(u,\ \beta_o)\, d\Lambda_o(u)$. Then (25) is $D_i^3(g_2, g_{3^*}) = C_1(g_2) - \pi_{io}^{-1}(R_i - \pi_{io}) g_{i3^*}$. For any fixed g_2, proposition 2.3 RRZ establishes that the g_{3^*} minimizing the variance of $D_i^3(g_2, g_{3^*})$ is, $g_{i,3}^{eff}(g_2) = E\Big[C_i(g_2)\Big|W_i\Big] = E\Big[D_i^{F3}\Big|W\Big]$. Hence the variance of $D_i^3(g_2, g_{3^*})$ is minimized by finding the g_2 that minimizes

$$var\, C_{i,1}(g_2^*) - 2\, cov\Big[C_{i,1}(g_2^*), \Big(\pi_{io}^{-1}(R_i - \pi_{io})\, g_{i,3}^{eff}\Big)\Big] \qquad (B.1)$$

Taking the expectations in (B.1) conditional on W_i, the only term containing g_2 is

$$E\Big[\pi_{io}^{-1}\,(1-\pi_{io})\,(i^{-1} g_{i,2}\, k_1)^2\Big]$$

which is minimized by $g_{i,2} = 0$.

Appendix C

In this appendix we show that the variance of $D_i^b(W^l)$ (27) is minimized iff $q^b S_i^l = \pi_{io}^{-1}$ $(R_i - \pi_{io})\, E\left[D_i^{Fb}\,|W_i\right]$. For any given set of scores, S_i^l, the variance is minimized when q^b $= P^{bl}$ (28). Since the variance is non-increasing in the dimension of S_i^l, $P^{bl} S_i^l = \pi_{io}^{-1}(R_i - \pi_{io}) g_b^{eff}$ iff, for all $W_i^m > W_i^l$,

$$E\left[\left(D_i^b(\pi_{i,o}) \; - P^{bl} S^l\right)^T S^{m|l}\right] = 0\,. \tag{C.1}$$

Here $S_i^{m|l}$ are the linearly independent matrix of scores from the residual of the projection of S_i^m on S_i^l. Taking the expectation of (C.1) conditional on H_i, and using MAR restriction (14), this becomes $E\left[\; W_i^{m|l}(1 - \pi_{io}) \;\right] E\left[\; D_i^{Fb} - \pi_{io}\, P^{bl}\, W_i^l \,\middle|\, W_i\right] = 0$, which is true *iff* $P^{bl}\, S_i^l = \pi_{io}^{-1}\,(R_i - \pi_{io})\, E\left[D_i^{Fb}\,|W_i\right]$.

Result 1: Taking each of the expectations in P^{bl} conditional on H_i, we obtain $P^{bl} = E$ $\left[(1 - \pi_{io}) D_i^{Fb} \times (W_i^l)^T\right] \times E\left[\pi_{io}(1 - \pi_{io})\, W_i^l W_i^{l^T}\right]^{-1}$. Let W^f have discrete covariate space of dimension f^*, with model (15) parameterized so that the design matrix is the $f^* \times f^*$ identity matrix. Then, the matrix of scores are orthonormal, and $P^{bf} S_i^f = \pi_{io}^{-1}(R_i - \pi_{io})$ $E\left[D_i^{Fb}\middle|W^f\right]$.

Result 2: To see that for (30) $P^{bf}\, S_i^f = \pi_{io}^{-1}\,(R_i - \pi_{io})\, g_b^{eff}$, note that (30) is correctly specified with $\psi_1 = 1$, $\psi_2 = 0$; by the general form of the P^{bl} in Result 1, the projection of $D_i^b(\pi_{i,o})$ onto $\pi_{io}^{-1}(R_i - \pi_{io})\, W_{ib}^{eff}$ is $\pi_{io}^{-1}(R_i - \pi_{io}) E\left[D_i^{Fb}\,\middle|\, W_i\right]$. Since the span of the scores from model (30) is greater than the span of $\pi_{io}^{-1}(R_i - \pi_{io}) W_i^{eff}$, (29) is true for the scores from model (30).

Appendix D

In this appendix we derive a general expression for $K_{v^\dagger}(W_i)$, and the specific expression given in (39). We define

$$\lambda_m^*\,(s|\, v^\dagger, A_i\,) = \lim_{h \to 0} \; Pr(\, s \le X_i < s + h, \Delta_i = m\,|\;\; X_i \ge s, v^\dagger, A_{i,}\,)\Big/ h\;; \tag{D.1}$$

$$\lambda_X^*\,(s|\, v^\dagger, A_i\,) = \sum_{m=0}^{1} \lambda_m^*(s\,|\, v^\dagger, A) \tag{D.2}$$

$$rr_{_m}(s \mid v^\dagger, A_i) = \lambda_{\,m}(s|\, v^\dagger, A_i\,)\Big/ \lambda_m\,(s|\, v^1, A_i,) \tag{D.3}$$

where $m \in \{0, 1\}$. We further define a *" non-informative censoring"* assumption

$$Pr(C_i \geq s|T_i \geq s, V_i, A_i) = Pr(C_i \geq s|T_i \geq s, A_i) \qquad (D.4)$$

The hazards in (D.1) are referred to as the crude hazards of C_i, $(m = 0)$, and T_i, $(m = 1)$. (D.2) is the hazard for the random variable X_i. Under the conditional independence assumption, the crude hazards equal the net hazards (e.g. Andersen et al., 1991). Non-informative censoring assumption (D.4) is similar in subject matter content to the usual non-informative censoring assumption, but does not imply that assumption: the latter allows for dependency of both C_i and T_i on V_i, but requires that, in terms of factorability of the likelihood, such dependency be distinct (Andersen et al., 1991).

By Bayes' rule, $Pr(v^\dagger|X_i, \Delta_i,\ A_i) = Pr(X_i, \Delta_i|v^\dagger,\ A_i) \times Pr(v^\dagger|\ A_i)\Big/Pr(X_i, \Delta_i|\ A_i)$. Writing $Pr(X_i = x_i, \Delta_i|\ v, A_i)$ as $\lambda_{\Delta_i}\Big(x_i\ |v, A_i, X_i \geq x_i\Big) \times Pr(X_i \geq x_i|v, A_i)$, we obtain

$$K_{i,v^\dagger} = rr_{\Delta_i}(x_i|\ v^\dagger, A_i,) \times Pr(X_i \geq x_i|v^\dagger, A_i)/\ Pr(X_i \geq x_i|v^1, A_i) \times \qquad (D.5)$$
$$Pr(v^\dagger|\ A_i)/Pr(v^1|\ A_i)$$

Applying (D.4), the right hand side of (D.5) gives

$$rr_1(x_i|\ v^\dagger, A_i)^{\Delta_i} \times Pr(T_i \geq x_i|v^\dagger, A_i)\big/Pr(T_i \geq x_i|v^1, A_i) \times \qquad (D.6)$$
$$Pr(v^\dagger|\ A_i)\big/Pr(v^1|\ A_i)$$

Using LE estimators as in section 9, one could postulate models for the distribution of T_i conditional on $\{V_i, A_i\}$ and estimate (D.6); with additional assumptions about the conditional distribution of C_i, (D.5) can be similarly estimated. When J contains all the independent risk factors in A_i, $(D.6)$ becomes (39).

Though (D.5-D.6, 38) are true regardless of the support of V_i and A_i, consistency of the estimators given in section 9 depends on the discreteness of A_i. When the support is not discrete, $K_{i,v}$ can be approximated by forming discretized random variables A_i^s, and using the empirical distribution of $Pr(V_i|A_i^s)$ instead of $Pr(V_i|A_i)$.

Table 1.

Estimating Survival Conditional on Hp and Age at 5.25 years in the Linxian Cohort

	Survival (95% CI)	
	***H. Pylori*- (V0)**	***H. Pylori*+ (V1)**
Young (J0)	**99.2 (98.9, 99.5)**	**98.8 (98.4, 99.0)**
Old (J1)	**97.3 (96.1, 98.1)**	**95.5 (94.4, 96.3)**

The estimates are based on the CPH model with relative risk exp(β_1V+ β_2 J).
The $\hat{\pi}(\Delta, J)$ -estimator (17) was used for estimating β_o and $\Lambda_o\left(\tau, \beta_o\right)$.

Table 2.

The STP, RC-efficient, and $\hat{\pi}(\Delta, J)$ Semiparametric Estimators of $S(\tau \mid v, j)$

	$S(\tau \mid v0, j0) = 90\%$			$S(\tau \mid v1, j1) = 73.5\%$		
Estimator	**Mean Survival V=0, J=0**	**95% CI Coverage**	**Relative Efficiency**	**Mean Survival V=1, J=1**	**95% CI Coverage**	**Relative Efficiency**
STP	**95.0**	**94.7**	**100**	**73.5**	**95.5**	**100**
RC-efficient	**95.0**	**95.0**	**55**	**73.6**	**95.7**	**27**
$\hat{\pi}(\Delta, J)$	**95.0**	**95.6**	**57**	**73.5**	**95.2**	**31**

Note: Relative efficiency equals 100 times the ratio of the variance of the estimator to the variance of the STP estimator.

Table 3.
The STP, $\hat{\pi}(\Delta, J)$, SLE, and ILE non-parametric estimators of $S(\tau \mid v)$ in the presence of an auxiliary covariate

Estimator	$S(\tau \mid v0) = 90\%$			$S(\tau \mid v1) = 81\%$		
	Mean Survival V=0	95% CI Coverage	Relative Efficiency	Mean Survival V=1	95% CI Coverage	Relative Efficiency
STP	90.0	94.1	100	81.0	94.7	100
$\hat{\pi}(\Delta)$	90.0	94.5	83	81.0	95.4	50
$\hat{\pi}(\Delta, J)$	90.0	93.8	74	81.0	94.8	45
SLE correct	90.0	94.4	71	81.0	95.7	40
ILE correct	90.0	94.4	71	81.0	95.7	40
SLE prior	90.0	94.9	85	81.0	95.4	43
ILE prior	90.0	94.2	71	81.0	95.5	40
SLE null	90.0	94.4	74	81.0	95.7	40
ILE null	90.0	94.4	71	81.0	95.8	40

Note: Relative efficiency equals 100 times the ratio of the variance of the estimator to the variance of the STP estimator